SEMESTER GASAL 2023/2024
ES234422 – PEMODELAN DAN ANALITIKA PREDIKTIF
TUGAS GROUP PROJECT (TGP) #1
EKSPLORASI DAN PRAPROSES DATA

TGP #1 berkaitan dengan eskplorasi data (*Exploratory Data Analysis*/EDA) dan praproses data terhadap sebuah data mengenai kampanye pemasaran langsung (*direct marketing*) produk deposito bank berjangka dari sebuah institusi perbankan di Portugal. Kampanye dilakukan melalui panggilan telpon di mana seorang klien mungkin harus dihubungi lebih dari sekali. Kampanye melalui pemasaran langsung ini ditujukan untuk memprediksi apakah seorang klien yang dihubungi akan membeli produk deposito berjangka tersebut atau tidak (*subscribe*).

A. Deskripsi data

Data kampanye pemasaran langsung adalah data yang diperoleh dari bulan Mei 2008 hingga November 2010. Set data ini terdiri dari 45.211 baris, 16 atribut (variabel independen) dan 1 atribut target (variabel dependen/target).

 # Bank client data:

(1) **age** (numeric)
(2) **job**: type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","studen, "blue-collar","self-employed","retired","technician","services")
(3) **marital**: marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
(4) **education** (categorical: "unknown","secondary","primary","tertiary")
(5) **default**: has credit in default? (binary: "yes","no")
(6) **balance**: average yearly balance, in euros (numeric)
(7) **housing**: has housing loan? (binary: "yes","no")
(8) **loan**: has personal loan? (binary: "yes","no")

 # Attribute related with the last contact of the current campaign:

(9) **contact**: contact communication type (categorical: "unknown","telephone","cellular")
(10) **day**: last contact day of the month (numeric)
(11) **month**: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
(12) **duration**: last contact duration, in seconds (numeric)

 # other attributes:

(13) **campaign**: number of contacts performed during this campaign and for this client (numeric, includes last contact)
(14) **pdays**: number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
(15) **previous**: number of contacts performed before this campaign and for this client (numeric)
(16) **poutcome**: outcome of the previous marketing campaign (categorical: "unknown", "other", "failure", "success")

Output variable (desired target):

(17) **subscribe**: has the client *subscribed* a term deposit? (binary: "yes", "no")

B. Tugas

1. Lakukan eksplorasi data dari berbagai perspektif untuk memahami karakteristik data agar anda mempunyai persepsi yang baik terhadap data. Gambakan hasil eksplorasi dalam berbagai bentuk grafik/*chart* yang menurut anda paling sesuai untuk menggambarkan karateristik data secara

komprehensip dan mudah dipahami. Anda dapat mengggunakan library visualisasi data yang populer dalam python (**Seaborn**) yang dijelaskan dalam tutorial atau yang anda peroleh sendiri.

Analisis eksploratori yang harus dilakukan paling tidak meliputi analisis ***univariate*** (analisis dari setiap atribut) dan analisis ***bivariate*** (analisis korelasi) antara setiap pasangan atribut independen dan setiap pasangan atribut independen dengan atribut dependen (atribut kelas, yaitu ***subscribe***).

Khusus untuk analisis bivariate, lakukan analisis dalam hal apa kemungkinan klien akan membeli produk (subscribe) dilihat dari setiap atribut independen. Sebagai contoh, dilihat dari nilai atribut "age", apakah klien dengan rentang usia tertentu mempunyai kecenderungan untuk membeli produk? Contoh lain, jika dilihat dari nilai atribut "job", apakah klien dengan jenis job tertentu mempunyai kecenderungan untuk membli produk. Lakukan analisis serupa dilihat dari nilai atribut lainnya.

Selain itu, dalam analisis bivariate antar setiap pasang atribut perlu dilakukan untuk nilai korelasi antar sepasang atribut. Jika korelasinya tinggi salah satu dari atribut dapat diabaikan. Tetapi untuk melihat yang mana dari kedua atribut tersebut yang harus dihapus dapat dilihat dari nilai korelasi antara atribut independen dan atribut dependen.

2. Lakukan praproses data yang menurut anda diperlukan sebelum dilakukan proses klasifikasi Praproses antara lain meliputi pengecekan setiap tipe data atribut dan juga penyesuaian tipe data (jika tipe data yang ditentukan oleh python secara otomatis dianggap kurang tepat atau tidak dikenali oleh python yaitu tipe data *object*). Selan itu juga perlu dilakukan praproses terkait dengan *missing value* (kalau ada), data pencilan (*oulier*), transformasi data. Praproses data dapat didasarkan pada hasil eksplorasi data, dan atribut yang sebaiknya dihapus (misalnyu dari hasil korelasi antar pasangan atribut independen seperti dijelaskan dalam poin 1).

C. Laporan dan Batas Waktu

- Laporan ditulis pada kertas berukuran A4 dengan spasi tunggal. Laporan dalam format PDF diserahkan per kelompok dan diunggah melalui myITS Calssroom **paling lambat pada tanggal 25 Maret 2024 pukul 10:00 WIB** (hanya satu orang dari setiap kelompok yang harus menyerahkan laporan).
- Penilaian akan didasarkan pada beberapa kriteria: sistematika serta kelengkapan dan kejelasan termasuk tata tulis (20%), kejelasan metode dan uraian hasil eksplorasi data (40%), dan kejelasan metode dan uraian praproses data (40%).
- Selain laporan tugas (format PDF), setiap kelompok juga harus mengunggah *script program python* (s*aved file*) dalam format **.ipnyb*.
- **Isi laporan dan script program yang mengindikasikan adanya plagiarisme dari berbagaiu sumber TIDAK AKAN DINILAI**

----------oooOooo----------